Journal of Classification 34:49-75 (2017)
DOI: 10.1007/s00357-017-9226-x

Maximal Interaction Two-Mode Clustering

Jan Schepers

Maastricht University, The Netherlands and K.U. Leuven, Belgium

Hans-Hermann Bock

RWTH Aachen University, Germany

Iven Van Mechelen

K.U. Leuven, Belgium

Abstract: Most classical approaches for two-mode clustering of a data matrix are designed to attain homogeneous row by column clusters (blocks, biclusters), that is, biclusters with a small variation of data values within the blocks. In contrast, this article deals with methods that look for a biclustering with a large interaction between row and column clusters. Thereby an aggregated, condensed representation of the existing interaction structure is obtained, together with corresponding row and column clusters, which both allow a parsimonious visualization and interpretation. In this paper we provide a statistical justification, in terms of a probabilistic model, for a two-mode interaction clustering criterion that has been proposed by Bock (1980). Furthermore, we show that maximization of this criterion is equivalent to minimizing the classical least-squares two-mode partitioning criterion for the double-centered version of the data matrix. The latter implies that the interaction clustering criterion can be optimized by applying classical two-mode partitioning algorithms. We illustrate the usefulness of our approach for the case of an empirical data set from personality psychology and we compare this method with other biclustering approaches where interactions play a role.

Keywords: Two-mode data; Biclustering; Capturing row by column interaction; Clustering criteria; Probabilistic clustering model; Classification likelihood.

The research reported in this paper was supported by the Flemish Fund for Scientific Research (G.0546.09), by the Interuniversity Attraction Poles program financed by the Belgian government (IAP P7/06), and by the Research Fund of K.U. Leuven (GOA/2015/03).

Corresponding Author's Address: J. Schepers, Faculty of Psychology and Neuroscience, Maastricht University, 6200 MD Maastricht, The Netherlands, email: jan.schepers@maastrichtuniversity.nl.

Published online: 20 March 2017

1. Introduction

Observed data often can be represented under the form of an I by J data matrix $\mathbf{D} = (d_{i,j})$. We consider the case in which the rows and columns of the data matrix constitute the levels of two categorical variables X and Y, say, with I and J categories, respectively, and the cell entries denote the observed values of a single quantitative dependent variable d. In the terminology introduced by Carroll and Arabie (1980), this type of data is called two-way two-mode. Such data matrices are collected, for instance, in contextualized personality research, when a set of I individuals (labeled by $i = 1, ..., I$) is measured on some behavior of interest d in J different situations (labeled by $j = 1, ..., J$). Other examples can be found in the study of micro-array data in genome research where DNA expression levels $d_{i,j}$ are obtained for I genes under J different conditions, or in agricultural studies where, for instance, crop yield per hectare is recorded for crops of I different genotypes and at J different locations.

A major challenge in these and other fields of science is to capture the dominant structural information as included in data matrices. However, typically the size $I \times J$ of the data matrix $\mathbf{D}$ is large, by which we mean too large to succinctly describe and interpret the full information as included in the data at hand. Then understanding the overall structure of this information is a challenge. A useful way to resolve this problem then is to simultaneously cluster the rows and columns of the matrix $\mathbf{D}$ in P row clusters and Q column clusters, respectively, such that the $I \times J$ data matrix is partitioned into PQ biclusters (blocks) and its structural information is represented as much as possible in a condensed parsimonious way by PQ block-specific values.

Many two-mode clustering methods have been developed (for an overview, see Van Mechelen, Bock, and De Boeck, 2004; Govaert and Nadif, 2013) and, for a given data set and a substantive research question at hand, some of these are more suitable than others. We will focus on a two-mode clustering criterion first proposed by Bock (1980), which explicitly addresses the *row by column interaction* in $\mathbf{D}$ (see formula (2) below). After Bock (1980), various two-mode clustering methods were proposed that use interactions concepts. However, most of these methods look for (possibly overlapping) blocks in the data matrix with a minimal within-block row by column interaction rather than for a representation of the full row by column interaction in the data matrix $\mathbf{D}$. (These methods will be commented on in Section 6.1.) The criterion proposed by Bock (1980), which we will further call the *maximal interaction two-mode clustering* criterion, implies simultaneously looking for a partition of the row set $\mathcal{X} = \{1, ..., I\}$ and a partition of the column set $\mathcal{Y} = \{1, ..., J\}$ which are such that the implied interaction among the row clusters and the column clusters is maximal, on the average.

In this paper, we address two issues pertaining to maximal interaction clustering that have not been addressed so far: First, we develop a statistical justification for the criterion proposed by Bock and second, we show how the criterion can be optimized numerically. Concerning the first aspect we will describe a probabilistic model from which the *maximal interaction two-mode clustering* criterion results when using a classification likelihood approach for model estimation (Section 3). This is useful for understanding the conditions under which the proposed interaction criterion is likely to be successful (Banfield and Raftery, 1993; Bock, 1996) and helps clarifying its relation to other clustering methods. Concerning the second aspect we will show (Section 4) that optimizing the criterion in question is equivalent to minimizing a standard least-squares two-mode partitioning criterion for a suitably transformed version of the data matrix. The latter result implies that the optimal solution for maximal interaction two-mode clustering can be obtained by means of existing classical two-mode partitioning algorithms, some of which are available as free software and some of which have been tested extensively with regard to numerical performance (e.g., Van Rosmalen et al., 2009). In addition, we will apply the maximal interaction clustering approach to an empirical data set from personality research in psychology and show how it can capture indeed the gist of the interaction pattern of a data matrix under study (Section 5).

The remainder of this paper is organized as follows. In the next section, the maximal interaction two-mode clustering criterion is explained. Subsequently, in Section 3, a statistical justification is given for this criterion. Next, in Section 4, it is shown that maximal interaction two-mode clustering is equivalent to classical least-squares two-mode partitioning when applied to a suitable transformation of the data. Section 5 presents the data example and Section 6 provides a general discussion of our approach together with a detailed comparison to other interaction-based biclustering methods (Section 6.1). The paper ends with some concluding remarks.

2. Maximal Interaction Two-Mode Clustering

2.1 Motivating Research Problems

Although it is not possible to find a single clustering criterion that is uniformly better than all other possible criteria, one criterion can be better than another one for a specific research question. For instance, when performing a biclustering in the usual way (with a deterministic method such as double k-means, or with a stochastic model such as some two-mode mixture model, which may be estimated in a frequentist or a Bayesian way) the resulting biclustering will be dominated by the row and column main effects.

This results from the fact that these methods essentially rely on a clustering structure to capture the main effects and the row by column interaction as well (see Section 3.2). However, in many applications the focus of interest of the researchers is not so much in the main effects of the rows and columns, but more in the interaction between them (see also the references in Section 6.1). As we will explain in Section 2.2, maximal interaction clustering implies a simultaneous clustering of the rows and columns of the matrix **D** in P row clusters and Q column clusters, respectively, such that the row by column interaction pattern is highlighted as much as possible in a parsimonious way by PQ block-specific interaction values. This amounts to assuming that all interaction terms within the same bicluster are equal (whereas the main effects should play no role, see Section 3.1). In the remainder of this section we describe some applications where maximal interaction clustering is well suited.

In contextualized personality psychology, a critical challenge is to capture person by situation interactions (Mischel and Shoda, 1995, 1998; Geiser et al., 2015). Indeed, a key question addressed by researchers in this field is whether the situation effect is the same for all individuals and, if not, what the structure of the person by situation interaction looks like. Furthermore, contextualized personality psychologists are typically less interested in situation main effects. The latter are considered to be part of general psychology, are often trivial and correspond to common sense. For instance, it should not come as a surprise to find that people respond more aggressively in situations that are more frustrating. Contextualized personality psychologists are, however, primarily interested in the shape of the individual-specific *behavioral signature* (i.e., the response pattern across different situations), in which the global level of this profile (i.e., the subject main effect) is less important or even of no interest at all. The shape of the signature is considered to be an important characteristic of an individual and contextualized personality psychologists advocate that it constitutes an essential part to the study of personality (Shoda et al., 2013, 2015). For example, individuals might be characterized by different sensitivities to specific types of frustration such as responding aggressively as a result of being let down by others versus as a result of being narcissistically offended.

Another domain of application is agriculture where researchers obtain, from suitable field experiments, large genotype by environment data matrices on, for instance, crop yield with the research focus typically being on the genotype by environment interaction (G×E interaction: Corsten and Denis, 1990; Piepho, 1997, 1999), rather than on the genotype and environment main effects. For instance, the amount of rainfall may differ between locations and for plant breeders it is important to know whether in terms of crop yield genotypes are differentially sensitive to these variations across

locations. If so, this "seriously limits efforts in selecting superior genotypes for both new crop introduction and improved cultivar development" (Shafii and Price, 1998). Moreover, it is then important to *understand* this G×E interaction pattern in order to make region-specific recommendations with regard to choosing genotypes and/or selecting locations for optimal crop yield. State-of-the-art methods for analyzing data from G×E studies include AMMI (additive main effects and multiplicative interaction effects) models (Gollob, 1968; Gauch, 2006; Gauch, Piepho, and Annicchiarico, 2008; Forkman and Piepho, 2014), which assume individual row and column main effects but yield a parsimonious representation of the row by column interaction structure by decomposing the latter into a small set of principal components. The idea is that the row (genotype/plant) and column (location) main effects should not be taken into account when looking for a parsimonious representation of the data matrix because, likely, different genotypes show individual stress effects and individual deviation of sunshine/soil fertility may be specific to each location. In contrast, AMMI models wish to characterize plant types in terms of their *sensitivity* to, for instance, soil type (sandy, organic, etc.) or other environmental characteristics such as altitude, wind, and so on. Maximal interaction clustering bears a close resemblance to these methods in that it also looks for a parsimonious representation of the row by column interaction that is invariant to the magnitude of the row and column main effects. However, this parsimonious representation of the interaction structure is obtained by means of a two-mode clustering, which yields results that may be considered more easily interpretable. In this regard one may note that the usual procedure when applying AMMI models is to assume two or three components and then to use biplots (Gabriel, 1971; Gower and Hand, 1996) in order to identify 'clusters' of similar genotypes and of similar environments. In our approach, this clustering is implied by the very nature of the method, obviating the need for a two-step approach.

Obviously, apart from agriculture, gene by environment interactions are also a key topic of interest in medicine (including psychopatology and psychiatry) (see, e.g., Hunter, 2005; Caspi and Moffitt, 2006; Moffitt, Caspi, and Rutter, 2006; National Institute of Environmental Health Services, 2016). Other applications may be found, for example, in marketing research (consumer behavior under different advertisement strategies).

2.2 Method

Consider a data matrix $\mathbf{D} = (d_{i,j})_{I\times J}$ where $d_{i,j}$ denotes the observed value on some criterion variable d for level i of a first categorical predictor variable X and level j of a second categorical predictor variable Y. Let $\mathcal{R} = \{R_1, ..., R_p, ..., R_P\}$ and $\mathcal{C} = \{C_1, ..., C_q, ..., C_Q\}$ denote par-

titions of the row set $\mathcal{X}$ and the column set $\mathcal{Y}$, comprising P and Q clusters, respectively, and let $\#R_p$ and $\#C_q$ denote the cluster cardinalities of row cluster R_p and column cluster C_q, respectively. Furthermore, let the two-mode cluster (bicluster, block cluster) $R_p \times C_q = \{(i,j)|i \in R_p, j \in C_q\}$ denote the Cartesian product of row cluster R_p and column cluster C_q. The observed amount of interaction associated with $R_p \times C_q$ can then be represented by the block interaction term

$$g_{p,q} = \bar{d}_{R_p,C_q} - \bar{d}_{R_p,\cdot} - \bar{d}_{\cdot,C_q} + \bar{d}_{\cdot,\cdot}, \tag{1}$$

where here and in the following we will use the notation:

- $\bar{d}_{i,\cdot} = \frac{1}{J}\sum_{j=1}^{J} d_{i,j}$ and $\bar{d}_{\cdot,j} = \frac{1}{I}\sum_{i=1}^{I} d_{i,j}$ for the row and column means,
- $\bar{d}_{R_p,\cdot} = \frac{1}{\#R_p \cdot J}\sum_{i\in R_p}\sum_{j=1}^{J} d_{i,j} = \frac{1}{\#R_p}\sum_{i\in R_p}\bar{d}_{i,\cdot}$ for the mean value in row cluster R_p,
- $\bar{d}_{\cdot,C_q} = \frac{1}{I\cdot\#C_q}\sum_{i=1}^{I}\sum_{j\in C_q} d_{i,j} = \frac{1}{\#C_q}\sum_{j\in C_q}\bar{d}_{\cdot,j}$ for the mean value in column cluster C_q,
- $\bar{d}_{R_p,C_q} = \frac{1}{\#R_p\cdot\#C_q}\sum_{i\in R_p}\sum_{j\in C_q} d_{i,j}$ for the mean value in block $R_p \times C_q$,
- $\bar{d}_{\cdot,\cdot} = \frac{1}{I\cdot J}\sum_{i=1}^{I}\sum_{j=1}^{J} d_{i,j}$ for the overall mean in $\mathbf{D}$.

Bock (1980) proposed to look for a two-mode partitioning (biclustering) $\mathcal{R}\times\mathcal{C} = \{R_p \times C_q; p = 1,...,P, q = 1,...,Q\}$, for given numbers of clusters P and Q, that maximizes the overall interaction criterion

$$f(\mathcal{R},\mathcal{C}) := f(\mathcal{R},\mathcal{C};\mathbf{D}) := \sum_{p=1}^{P}\sum_{q=1}^{Q} \#R_p \cdot \#C_q \cdot g_{p,q}^2. \tag{2}$$

Two concerns are warranted with regard to criterion (2). Firstly, one may wonder to what degree this criterion can be justified in terms of a probabilistic model for the data $d_{i,j}$. In the next section, this question will be answered by showing that an insightful probabilistic ANOVA model leads to the criterion in question. Secondly, maximizing (2) over all possible combinations of row and column partitions is a challenging combinatorial optimization problem for which no analytical solution exists and for which a complete

enumeration of all possible solutions is computationally infeasible unless the number of rows and columns of **D** is very small. Therefore, in order to apply *maximal interaction two-mode clustering* to large data matrices, suitable approximate numerical optimization algorithms are needed. In Section 4, it will be shown that the problem of maximizing (2) is equivalent to minimizing a classical least-squares two-mode partitioning criterion for the double-centered data matrix. This fact essentially resolves the optimization problem for (2) since there exist a range of good numerical algorithms designed for classical least-squares two-mode partitioning, which provides the possibility of analyzing large empirical data sets by means of the maximal interaction two-mode clustering approach.

3. Statistical Justification

3.1 A Probabilistic Model for Maximal Interaction Two-Mode Clustering

In this section, a statistical justification for interaction criterion (2) is given in terms of a probabilistic ANOVA model for the data. In particular, we will show in this section that maximizing the classification likelihood for this model is equivalent to maximizing interaction criterion (2).

The model in question describes a situation in which each row i and each column j has its specific additive main effect (α_i and β_j, respectively), but the interaction terms are the same for all (i, j) within block $R_p \times C_q$:

$$d_{i,j} = \mu + \alpha_i + \beta_j + \gamma_{p,q} + \epsilon_{i,j} \qquad i \in R_p, j \in C_q, p = 1, ..., P, q = 1, ..., Q. \tag{3}$$

The error terms $\epsilon_{i,j}$ are $iid\ \mathcal{N}(0, \sigma^2)$ variables, and the main effects and block-specific interaction terms are identified as follows:

$$\bar{\alpha}_\cdot = \frac{1}{I}\sum_{i=1}^{I} \alpha_i = 0 \tag{4a}$$

$$\bar{\beta}_\cdot = \frac{1}{J}\sum_{j=1}^{J} \beta_j = 0 \tag{4b}$$

$$\bar{\gamma}_{p,\cdot} = \frac{1}{J}\sum_{q=1}^{Q} \#C_q \cdot \gamma_{p,q} = 0 \tag{4c}$$

$$\bar{\gamma}_{\cdot,q} = \frac{1}{I}\sum_{p=1}^{P} \#R_p \cdot \gamma_{p,q} = 0. \tag{4d}$$

Let $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_I)$, $\boldsymbol{\beta} = (\beta_1, ..., \beta_J)$, $\boldsymbol{\gamma} = (\gamma_{1,1}, ..., \gamma_{p,q})$, and $\boldsymbol{\theta} = (\mu, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$. The classification likelihood (McLachlan, 1982) or fixed-

classification (Bock, 1996) estimation of model (3) proceeds by maximizing the likelihood over all possible two-mode partitions $\mathcal{R} \times \mathcal{C}$ and all possible values of the parameters μ, α_i, β_j and $\gamma_{p,q}$, that is:

$$L(\mathcal{R},\mathcal{C},\boldsymbol{\theta}) = \prod_{p=1}^{P}\prod_{q=1}^{Q}\prod_{i\in R_p}\prod_{j\in C_q}\frac{1}{\sigma}\exp\Big(-\frac{1}{2}\frac{|d_{i,j}-\mu-\alpha_i-\beta_j-\gamma_{p,q}|^2}{\sigma^2}\Big), \tag{5}$$

subject to identification constraints (4a-4d) for given numbers of clusters P and Q.

For a given two-mode partitioning $\mathcal{R} \times \mathcal{C}$, maximum likelihood (m.l.) estimation of the unknown parameters μ, α_i, β_j, and $\gamma_{p,q}$ amounts to minimizing the quadratic criterion:

$$S := \sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{i\in R_p}\sum_{j\in C_q}|d_{i,j}-\mu-\alpha_i-\beta_j-\gamma_{p,q}|^2, \tag{6}$$

subject to identification constraints (4a-4d). This is obvious for the case of a known variance σ^2, but also holds for the case of an unknown σ^2 (where the maximum of (5) will typically be $+\infty$, such that we have to restrain to a local maximum w.r.t. $\sigma^2 > 0$).

At this point, it is convenient to introduce the following statistics:

- $\tilde{\mu} = \bar{d}_{\cdot,\cdot}$,
- $\tilde{\alpha}_i = \bar{d}_{i,\cdot} - \bar{d}_{\cdot,\cdot}$,
- $\tilde{\beta}_j = \bar{d}_{\cdot,j} - \bar{d}_{\cdot,\cdot}$,
- $\tilde{\gamma}_{p,q} = \bar{d}_{R_p,C_q} - \bar{d}_{R_p,\cdot} - \bar{d}_{\cdot,C_q} + \bar{d}_{\cdot,\cdot} = g_{p,q}$

for $i = 1,...,I, j = 1,...,J, p = 1,...,P$ and $q = 1,...,Q$.

Proposition 1. *For the ANOVA model (3) the m.l. estimates of the (standardized) main and interaction effects are given by*

$$\hat{\mu} = \tilde{\mu}, \qquad \hat{\alpha}_i = \tilde{\alpha}_i, \qquad \hat{\beta}_j = \tilde{\beta}_j, \qquad \textit{and} \qquad \hat{\gamma}_{p,q} = \tilde{\gamma}_{p,q}$$

for $i = 1,...,I, j = 1,...,J, p = 1,...,P$ *and* $q = 1,...,Q$.

Proof. Whereas this result could be derived from sufficiency concepts for exponential distribution families, we present here an elementary algebraic proof. After inserting $\tilde{\mu}$, $\tilde{\alpha}_i$, $\tilde{\beta}_j$, and $\tilde{\gamma}_{p,q}$ (which are to be treated as constants), the squared-error residual sum S can be decomposed as follows:

$$
\begin{aligned}
S &= \sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{i\in R_p}\sum_{j\in C_q} |(d_{i,j}-\tilde{\mu}-\tilde{\alpha}_i-\tilde{\beta}_j-\tilde{\gamma}_{p,q}) \\
&\quad + (\tilde{\mu}-\mu)+(\tilde{\alpha}_i-\alpha_i)+(\tilde{\beta}_j-\beta_j)+(\tilde{\gamma}_{p,q}-\gamma_{p,q})|^2 \\
&= \sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{i\in R_p}\sum_{j\in C_q} |d_{i,j}-\tilde{\mu}-\tilde{\alpha}_i-\tilde{\beta}_j-\tilde{\gamma}_{p,q}|^2 \\
&\quad + \sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{i\in R_p}\sum_{j\in C_q} [|\tilde{\mu}-\mu|^2+|\tilde{\alpha}_i-\alpha_i|^2 \\
&\quad + |\tilde{\beta}_j-\beta_j|^2+|\tilde{\gamma}_{p,q}-\gamma_{p,q}|^2] + 2\cdot U
\end{aligned}
$$

where U is a sum of cross-product terms that equals 0 (see below). Since the second sum is always non-negative, S is minimized if and only if $\mu=\tilde{\mu}$, $\alpha_i=\tilde{\alpha}_i$, $\beta_j=\tilde{\beta}_j$, and $\gamma_{p,q}=\tilde{\gamma}_{p,q}$ as asserted. For U we have

$$
\begin{aligned}
U = \sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{i\in R_p}\sum_{j\in C_q} & (d_{i,j}-\tilde{\mu}-\tilde{\alpha}_i-\tilde{\beta}_j-\tilde{\gamma}_{p,q}) \\
& \cdot [\underbrace{(\tilde{\mu}-\mu)}_{A_1}+\underbrace{(\tilde{\alpha}_i-\alpha_i)}_{A_2}+\underbrace{(\tilde{\beta}_j-\beta_j)}_{A_3}+\underbrace{(\tilde{\gamma}_{p,q}-\gamma_{p,q})}_{A_4}] \\
+ \sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{i\in R_p}\sum_{j\in C_q} & [(\tilde{\mu}-\mu)\cdot\{\underbrace{(\tilde{\alpha}_i-\alpha_i)}_{B_1}+\underbrace{(\tilde{\beta}_j-\beta_j)}_{B_2}+\underbrace{(\tilde{\gamma}_{p,q}-\gamma_{p,q})}_{B_3}\} \\
+(\tilde{\alpha}_i-\alpha_i)\cdot\{\underbrace{(\tilde{\beta}_j-\beta_j)}_{C_1}+\underbrace{(\tilde{\gamma}_{p,q}-\gamma_{p,q})}_{C_2}\} & + (\tilde{\beta}_j-\beta_j)\cdot\underbrace{(\tilde{\gamma}_{p,q}-\gamma_{p,q})}_{C_3}]
\end{aligned}
$$

and we will show that $U=0$. In fact, the sum of the deviations

$$
\begin{aligned}
\Delta_{i,j} &:= d_{i,j}-\tilde{\mu}-\tilde{\alpha}_i-\tilde{\beta}_j-\tilde{\gamma}_{p,q} && (7)\\
&= d_{i,j}-\bar{d}_{i,\cdot}-\bar{d}_{\cdot,j}-\bar{d}_{R_p,C_q}+\bar{d}_{R_p,\cdot}+\bar{d}_{\cdot,C_q} && (8)
\end{aligned}
$$

over $i\in R_p, j\in C_q$ equals 0 since

$$
\begin{aligned}
&\sum_{i\in R_p}\sum_{j\in C_q}\Delta_{i,j} \\
&\quad \overset{(7)}{=} \sum_{i\in R_p}\sum_{j\in C_q}(d_{i,j}-\tilde{\mu}-\tilde{\alpha}_i-\tilde{\beta}_j)-\#R_p\cdot\#C_q\cdot\tilde{\gamma}_{p,q} && (9)\\
&\quad \overset{(8)}{=} \#R_p\cdot\#C_q\cdot(\bar{d}_{R_p,C_q}-\bar{d}_{R_p,\cdot}- \\
&\qquad\qquad \bar{d}_{\cdot,C_q}-\bar{d}_{R_p,C_q}+\bar{d}_{R_p,\cdot}+\bar{d}_{\cdot,C_q}) \\
&\quad = 0.
\end{aligned}
$$

Therefore, in U the partial sums related to A_1 and A_4 are 0 as well. Because of identification constraint (4a) and the fact that by definition $\bar{\tilde{\alpha}}_{\cdot} = 0$, the partial sum related to B_1 is 0 as well. The same holds for the partial sums related to B_2 and B_3, respectively, after considering the identification constraints (4b) and (4c-4d), and the definitions of $\tilde{\beta}_j$ and $\tilde{\gamma}_{p,q}$. The sum related to C_1 is 0 because of identification constraint (4b) and the fact that by definition $\bar{\tilde{\beta}}_{\cdot} = 0$. The sum related to C_2 is 0 because of identification constraint (4c) and the fact that by definition $\bar{\tilde{\gamma}}_{p,\cdot} = 0$. Similarly, the sum related to C_3 vanishes because of identification constraint (4d) and the fact that by definition $\bar{\tilde{\gamma}}_{\cdot,q} = 0$. Finally, the sum related to A_2 is, considering the fact that by definition $\bar{\tilde{\beta}}_{\cdot} = 0$ and $\bar{\tilde{\gamma}}_{p,\cdot} = 0$, given by

$$\begin{aligned}
&\textstyle\sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{i\in R_p}\sum_{j\in C_q}(d_{i,j}-\tilde{\mu}-\tilde{\alpha}_i-\tilde{\beta}_j-\tilde{\gamma}_{p,q})\cdot(\tilde{\alpha}_i-\alpha_i)\\
&\quad= J\cdot\textstyle\sum_{p=1}^{P}\sum_{i\in R_p}(\bar{d}_{i,\cdot}-\tilde{\mu}-\tilde{\alpha}_i-\bar{\tilde{\beta}}_{\cdot}-\bar{\tilde{\gamma}}_{p,\cdot})\cdot(\tilde{\alpha}_i-\alpha_i)\\
&\quad= J\cdot\textstyle\sum_{p=1}^{P}\sum_{i\in R_p}(\bar{d}_{i,\cdot}-\tilde{\mu}-\tilde{\alpha}_i-0-0)\cdot(\tilde{\alpha}_i-\alpha_i)\\
&\quad= J\cdot\textstyle\sum_{p=1}^{P}\sum_{i\in R_p}(\bar{d}_{i,\cdot}-\bar{d}_{\cdot,\cdot}-\bar{d}_{i,\cdot}+\bar{d}_{\cdot,\cdot})\cdot(\tilde{\alpha}_i-\alpha_i)=0.
\end{aligned}$$

In a similar way we show that the sum related to A_3 is 0 as well. So U is a sum of zero sums and equals 0.

■

It remains to consider optimization with respect to the two-mode partitioning $\mathcal{R}\times\mathcal{C}$.

Proposition 2. *Maximizing likelihood criterion (5), or minimizing quadratic criterion (6), is equivalent to maximizing interaction criterion (2).*

Proof. Substituting the m.l. estimates of μ, α_i, β_j and $\gamma_{p,q}$ into equation (6) for S, we obtain the following quadratic criterion, which is to be minimized with respect to $\mathcal{R}$ and $\mathcal{C}$:

$$\begin{aligned}
\tilde{S} &:= \sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{i\in R_p}\sum_{j\in C_q}|d_{i,j}-\hat{\mu}-\hat{\alpha}_i-\hat{\beta}_j-\hat{\gamma}_{p,q}|^2\\
&= \sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{i\in R_p}\sum_{j\in C_q}[|d_{i,j}-\hat{\mu}-\hat{\alpha}_i-\hat{\beta}_j|^2+\hat{\gamma}_{p,q}^2\\
&\quad -2\cdot(d_{i,j}-\hat{\mu}-\hat{\alpha}_i-\hat{\beta}_j)\cdot\hat{\gamma}_{p,q}]\\
&= const.+\sum_{p=1}^{P}\sum_{q=1}^{Q}\#R_p\cdot\#C_q\cdot\hat{\gamma}_{p,q}^2\\
&\quad -2\cdot\sum_{p=1}^{P}\sum_{q=1}^{Q}\hat{\gamma}_{p,q}\sum_{i\in R_p}\sum_{j\in C_q}(d_{i,j}-\hat{\mu}-\hat{\alpha}_i-\hat{\beta}_j)
\end{aligned}$$

$$\begin{aligned} &\stackrel{(9)}{=} const. + \sum_{p=1}^{P}\sum_{q=1}^{Q} \#R_p \cdot \#C_q \cdot \hat{\gamma}_{p,q}^2 \\ &\quad -2 \cdot \sum_{p=1}^{P}\sum_{q=1}^{Q} \hat{\gamma}_{p,q} \cdot (\#R_p \cdot \#C_q \cdot \hat{\gamma}_{p,q}) \\ &= const. - 1 \cdot \sum_{p=1}^{P}\sum_{q=1}^{Q} \#R_p \cdot \#C_q \cdot \hat{\gamma}_{p,q}^2. \end{aligned}$$

Since $\hat{\gamma}_{p,q} = g_{p,q}$ as stated by Proposition 1, minimizing $\tilde{S}$ (i.e., maximizing likelihood function (5)) over all $\mathcal{R} \times \mathcal{C}$ is equivalent to maximizing interaction criterion (2).
■

Remark 1: The previous model, formulas, and results apply also to the case when the data $d_{i,j}$ are multi-dimensional with values in $\mathbb{R}^k$, say, with i.i.d. normal errors $\epsilon_{i,j} \sim \mathcal{N}_k(0, \sigma^2 \mathbf{I}_k)$. The only change consists in replacing the absolute values $|...|$ by the Euclidean norm $||...||$ and dot products by the inner product in $\mathbb{R}^k$.

3.2 An ANOVA Model with Clustered Main Effects

Model (3) comprises $I + J$ individual main effects and $P \cdot Q$ block-specific interaction effects. An even more parsimonious model would read as follows:

$$d_{i,j} \quad = \quad \mu + \alpha_p + \beta_q + \gamma_{p,q} + \epsilon_{i,j} = \mu_{p,q} + \epsilon_{i,j}, \tag{10}$$

(for $i \in R_p, j \in C_q, p = 1, ..., P, q = 1, ..., Q$), where α_p and β_q represent $P + Q$ *cluster-specific* main effects (with suitable identification constraints) of row cluster R_p and column cluster C_q, respectively (instead of the $I + J$ main effects in the former model (3)). For this model, maximizing the classification likelihood comes down to minimizing the quadratic criterion

$$h(\mathcal{R}, \mathcal{C}; \mathbf{M}) := \sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{i \in R_p}\sum_{j \in C_q} |d_{i,j} - \mu_{p,q}|^2,$$

with respect to $\mathcal{R}$, $\mathcal{C}$ and the block centroid matrix $\mathbf{M} = (\mu_{p,q})_{P \times Q}$. Since, for a given two-mode partitioning $\mathcal{R} \times \mathcal{C}$, partial optimization w.r.t. $\mathbf{M}$ yields the block means (m.l. estimates)

$$\hat{\mu}_{p,q} = \bar{d}_{R_p, C_q} \qquad \text{for } p = 1, ..., P, q = 1, ..., Q,$$

this biclustering problem implies minimizing the criterion

$$H(\mathcal{R},\mathcal{C};\mathbf{D}) := h(\mathcal{R},\mathcal{C},\hat{\mathbf{M}}) := \sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{i\in R_p}\sum_{j\in C_q}|d_{i,j}-\bar{d}_{R_p,C_q}|^2, \quad (11)$$

with respect to $\mathcal{R}$ and $\mathcal{C}$, which is the classical least-squares two-mode partitioning problem (Van Mechelen, Bock, and De Boeck, 2004, pp. 373–374), sometimes referred to as double k-means.

Comparison of models (3) and (10) hence clarifies that maximizing interaction criterion (2) means concentrating on the row by column interaction only (while main effects may be row- and column-specific and insofar without any clustering structure), whereas, minimizing the classical least-squares two-mode partitioning criterion (11) tacitly assumes a clustering structure for the main effects as well, with the same clusters as for the row by column interaction.

4. Equivalence of Maximal Interaction Two-Mode Clustering and Least-Squares Two-Mode Partitioning of Double-Centered Data

In this section, it will first be shown (Section 4.1) that any two-mode partitioning $\mathcal{R}\times\mathcal{C}$ that maximizes interaction criterion (2) minimizes the classical least-squares two-mode (double k-means) partitioning criterion (11) when applied to the data matrix after double-centering and vice versa. Second (Section 4.2), an important consequence of the equivalence relation proven in this section will be discussed, that is, that no new numerical optimization algorithms have to be developed for maximal interaction two-mode clustering.

4.1 Proof of Equivalence

Let $\mathbf{D}^* = (d^*_{i,j})$ denote the double-centered data matrix where

$$d^*_{i,j} = d_{i,j} - \bar{d}_{i,\cdot} - \bar{d}_{\cdot,j} + \bar{d}_{\cdot,\cdot} \quad (12)$$

is the *individual* deviation from additivity for (i,j).

Proposition 3. *Maximizing interaction criterion (2) w.r.t. the two-mode partitioning $\mathcal{R}\times\mathcal{C}$ is equivalent to minimizing the classical least-squares two-mode partitioning criterion (11) for the double-centered data matrix $\boldsymbol{D}^*$:*

$$H(\mathcal{R},\mathcal{C};\boldsymbol{D}^*) = \sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{i\in R_p}\sum_{j\in C_q}|d^*_{i,j}-\bar{d}^*_{R_p,C_q}|^2 \to \min_{\mathcal{R},\mathcal{C}}. \quad (13)$$

Remark 2: Formulation (13) can be interpreted in the way that the maximal interaction criterion (2) looks for a two-mode partitioning $\mathcal{R} \times \mathcal{C}$ such that the individual deviations from additivity $d^*_{i,j}$ are, on the average, as homogeneous as possible within the blocks $R_p \times C_q$.

Proof. For any two-mode partitioning $\mathcal{R} \times \mathcal{C}$, the total sum of squares T in the double-centered data matrix $\mathbf{D}^*$ can be decomposed into a within-block and a between-block SSQ. Using the fact that by definition $\bar{d}^*_{\cdot,\cdot} = 0$), we obtain:

$$\begin{aligned} T &:= \sum_{i=1}^{I}\sum_{j=1}^{J}(d^*_{ij})^2 \\ &= \sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{i\in R_p}\sum_{j\in C_q}(d^*_{i,j} - \bar{d}^*_{R_p,C_q})^2 \\ &\quad + \sum_{p=1}^{P}\sum_{q=1}^{Q}\#R_p \cdot \#C_q \cdot (\bar{d}^*_{R_p,C_q})^2. \end{aligned} \tag{14}$$

Since T does not depend on $\mathcal{R}$ and $\mathcal{C}$ it appears that any two-mode partitioning $\mathcal{R} \times \mathcal{C}$ that minimizes the first term at the right-hand side of (14) also maximizes the second term in that expression, and vice versa. Obviously, the first term at the right-hand side of (14) is the classical least-squares two-mode partitioning criterion (11) when applied to the double-centered data. The proof will now be completed by showing that the second term at the right-hand side of (14) is identical to the interaction criterion (2). This is obvious if we can show that $\bar{d}^*_{R_p,C_q}$ equals the block-specific interaction value $g_{p,q}$ from (1). In fact, making use of the definitions of $\bar{d}^*_{R_p,C_q}$, $d^*_{i,j}$, $\bar{d}^*_{R_p,\cdot}$, $\bar{d}^*_{\cdot,C_q}$, and $g_{p,q}$ it follows that

$$\begin{aligned} &\bar{d}^*_{R_p,C_q} \\ &= \frac{1}{\#R_p \cdot \#C_q}\sum_{i\in R_p}\sum_{j\in C_q} d^*_{i,j} \\ &= \frac{1}{\#R_p \cdot \#C_q}\sum_{i\in R_p}\sum_{j\in C_q}(d_{i,j} - \bar{d}_{i,\cdot} - \bar{d}_{\cdot,j} + \bar{d}_{\cdot,\cdot}) \\ &= \frac{1}{\#R_p \cdot \#C_q}\sum_{i\in R_p}\sum_{j\in C_q} d_{i,j} - \frac{1}{\#R_p}\sum_{i\in R_p}\bar{d}_{i,\cdot} - \frac{1}{\#C_q}\sum_{j\in C_q}\bar{d}_{\cdot,j} + \bar{d}_{\cdot,\cdot} \\ &= \bar{d}_{R_p,C_q} - \bar{d}_{R_p,\cdot} - \bar{d}_{\cdot,C_q} + \bar{d}_{\cdot,\cdot} \\ &= g_{p,q}. \end{aligned}$$

■

4.2 Important Algorithmic Consequence

Proposition 3 implies that any numerical optimization algorithm designed for classical least-squares two-mode partitioning, that is, designed for minimizing criterion (11), can also be used to maximize interaction criterion (2), just by substituting the original data matrix $\mathbf{D}$ in (11) by its double-centered version $\mathbf{D}^*$. Fortunately, many algorithms for minimizing (11) have been proposed and evaluated. A selection of work in this area may be found in Gaul and Schader (1996); Baier, Gaul, and Schader (1997); Hansohm (2001); Vichi (2001); Castillo and Trejos (2002); Rocci and Vichi (2008) and Van Rosmalen et al. (2009). Therefore, there is no need to propose and evaluate novel optimization algorithms designed specifically for maximizing interaction criterion (2).

5. Application to Altruism Data

In this section, maximal interaction two-mode clustering is illustrated using data from the domain of contextualized personality psychology (Mischel and Schoda, 1995, 1998; Shoda et al., 2013, 2015), which aims at characterizing individual differences in behavior profiles across situations. An important challenge in this regard is to capture the gist of the person by situation interaction as included in behavioral data. Studying such interactions may reveal the underlying mechanisms through which the behavior under study comes about.

5.1 Maximal Interaction Results for Altruism Data

The data of our application stem from a study on altruism (Quintiens, 1999). The key question that goes with these data is to retrieve the mechanisms underlying individual differences in helping behavior. A group of $I = 102$ persons was presented with a set of $J = 16$ vignettes, each of which described in a few sentences an emergency situation that typically occurs in the everyday life of students, with a victim that could possibly be helped by the participant. Two (abbreviated) examples of such situation descriptions are: 'In a very crowded grocery store you see a little boy, weeping and crying for his mum', and 'Your neighbors ask you to care for their pets while they are abroad during the summer holidays and in return allow you to make use of their swimming pool'. The persons were asked to rate each situation with respect to the extent to which they would be willing to help the victim in it. For this purpose they had to use a 7-point scale from 0 (*definitely not*) through 6 (*definitely yes*).

To capture the dominant interaction pattern in the person by situation willingness to help data matrix $\mathbf{D} = (d_{i,j})$, we used the maximal interaction two-mode clustering approach from Section 2.2 by minimizing the classical least-squares two-mode partitioning criterion (11) for the double-centered matrix $\mathbf{D}^*$ with an algorithm implemented in free and user-friendly software called *TwoMP* (Schepers and Hofmans, 2009). Given pre-specified values P and Q, this algorithm starts from some initial two-mode partitioning $\mathcal{R}^0 \times \mathcal{C}^0$ and proceeds by an alternating least-squares optimization approach in which the classifications of the row and column sets are updated in turn. Each update implies that, consecutively, rows (resp. columns) are optimally (re)assigned to one of the row (resp. column) clusters. At each evaluation of a candidate assignment, a corresponding update of the centroid matrix $\hat{\mathbf{M}} = (\hat{\mu}_{p,q}) = (\bar{d}_{R_p,C_q})$ is computed. The alternating steps of updating the classifications of the row and column sets, respectively, are continued until the value of criterion (11) no longer decreases (or, equivalently, criterion (2) no longer increases). This algorithm is guaranteed to find a locally optimal solution which, as is well-known, is not necessarily the global optimum. Therefore, *TwoMP* allows the user to specify a desired number of different runs of the algorithm, each of which is initialized by an independently generated random start, and, among the multiple estimated solutions, selects the one for which (11) takes a minimal value. One may note that this alternating least-squares algorithm is a special case of the so-called DRIFT algorithm which can also handle three-mode partitioning and which has been tested extensively with regard to algorithmic performance (Schepers, Van Mechelen, and Ceulemans, 2006).

The person by situation willingness to help data matrix was clustered for all combinations of numbers of clusters $P = 2, ..., 6$ and $Q = 2, ..., 6$, and with 100 random starts for each such combination. In the framework of two-mode partitioning problems, various procedures for choosing the appropriate numbers of clusters P and Q have been proposed in the literature (see, e.g., Ceulemans and Kiers, 2006; Schepers, Ceulemans, and Van Mechelen, 2008; and Wilderjans, Ceulemans, and Meers, 2013). Instead of presenting a full analysis of our data, we refer, for illustration purposes, only to the optimal biclusterings with $P+Q \leq 5$. For this subset of solutions, the best one (i.e., with maximal criterion value (2)) comprises $P = 3$ person clusters and $Q = 2$ situation clusters. The maximized value of interaction criterion (2) equals 223.9 for this solution. As the individual person by situation interaction sum of squares (14) equals $T = 1941.9$, this implies that by applying maximal interaction clustering 11.53% of this sum of squares is captured by the optimal biclustering with $P = 3$ person clusters and $Q = 2$ situation clusters.

5.2 Comparing to Results from Double K-Means and Two-Mode Mixture Results

We also analyzed the person by situation altruism data matrix $\mathbf{D} = (d_{i,j})$ by two other widely used two-mode clustering methods in the social sciences, double k-means (Vichi, 2001) and two-mode Gaussian mixture analysis (Govaert and Nadif, 2013), in order to compare the resulting solutions with the one we obtained using maximal interaction clustering.

The double k-means solution was obtained by minimizing the classical least squares criterion (11) with the software *TwoMP* whereas the two-mode Gaussian mixture solution was obtained using the R package *blockcluster* (Iovleff and Singh Bhatia, 2015). In order to maintain comparability with the results of maximal interaction clustering these two solutions were likewise obtained for $P = 3$ person clusters, $Q = 2$ situation clusters and making use of 100 different starts. Furthermore, for the Gaussian mixture solution, person and situation cluster labels were obtained by assigning each person (resp. situation) to the person (resp. situation) cluster for which its posterior cluster membership probability was the largest.

For the double k-means solution the value of interaction criterion (2) turned out to equal 139.32 (explaining 7.2% of T) and for the two-mode mixture solution the value of this criterion was 129.67 (explaining 6.7% of T). Hence, both double k-means and two-mode mixture analysis capture some of the person by situation interaction sum of squares. However, both methods perform worse in this regard than maximal interaction clustering (which explained 11.53% of T).

To measure the agreement between the different biclusterings obtained for the altruism data, we also calculated the Hubert and Arabie adjusted Rand indices (ARI: Hubert and Arabie, 1985) between the person (resp. situation) clusterings as obtained by each pair of methods. The situation clusterings as obtained from double k-means and two-mode Gaussian mixture are identical ($ARI = 1.00$), but the situation clustering as obtained from the maximal interaction biclustering is different from these ($ARI = .52$). Similarly, the person clusterings as obtained from double k-means and two-mode Gaussian mixture analysis are more similar to each other ($ARI = .53$) than to the person clustering as obtained from the maximal interaction method ($ARI = .18$ and $ARI = .13$). Note that, based on an extensive simulation study, Steinley (2004) concluded that values of ARI below .65 reflect poor agreement. For the altruism data this then implies that the maximal interaction clustering method yielded a biclustering that is substantively very different from the solutions yielded by double k-means and two-mode Gaussian mixture analysis, respectively.

5.3 Substantive Interpretation of Maximal Interaction Biclusters

In this section, we will discuss in detail some of the substantive implications and interpretations that are implied by the obtained maximal interaction biclustering solution. Table 1 and Figure 1 show the interaction terms $(g_{p,q})$ of all $2 \cdot 3 = 6$ pairs of person and situation clusters (block clusters). It appears that the largest block-specific interaction terms (i.e., deviations from additivity) are observed for person clusters PC_2 (20 persons) and PC_3 (34 persons) in situation cluster SC_2 (which includes 3 situations). Moreover, since the interaction terms associated to person cluster PC_1 are close to zero, it appears that for all persons within the person cluster PC_1 (48 persons), helping behavior across situations appears to be described accurately by an additive effect of the persons and situations.

To obtain a substantive psychological interpretation for the obtained row (person) and column (situation) clusters and to highlight the relevance of the obtained biclustering, we have also used additional analyses. In a first additional analysis, we have compared the situation clusters with an external rating from expert judges of the extent to which each situation j describes an equivocal event, that is, an event that can be interpreted in different ways. Formally, we have introduced the indicator variable V for membership of each situation j in column (situation) cluster SC_2 such that $V_j = 1$ (0) if $j \in SC_2$ $(j \in SC_1)$. It appeared that V had a relatively high correlation to the expert rating $(r = 0.46)$ such that we may conclude that column cluster SC_2 comprizes more or less ambiguous situations.

In a second additional analysis, we looked for a substantive interpretation of the person clusters by analyzing their relationship to 16 external dispositional variables Z_l $(l = 1, ..., 16)$ that measure the general feelings and attitudes towards helping behavior for the persons $i = 1, ..., 102$ and that were recorded in the study. The row (person) clusters were described by the person cluster membership variable W such that for a person i we set $W_i = 1$ (resp. 2, 3) if i is in cluster PC_1 (resp. PC_2, PC_3). In this framework, we studied if and how the variable W relates to the external variables Z_l by considering a multinomial logistic regression model with W as dependent variable and Z_l as predictor variables (where $W_i = 1$, i.e., membership in person cluster PC_1, was chosen as reference category):

$$\log\Big(\frac{P(W_i = c+1)}{P(W_i = 1)}\Big) = \theta_{c0} + \sum_{l=1}^{16} \theta_{cl} \cdot Z_{il}$$
$$c = 1, 2, i = 1, ..., 102, l = 1, ..., 16.$$

A forward selection strategy identified two of the 16 dispositional variables as significant $(p < .05)$ predictors of W, namely Z_1 (i.e., the extent to which one feels satisfied when being helped by others) and Z_2 (i.e., the extent to

Table 1. Block-specific interaction terms $g_{p,q}$ for all pairs of person (PC_p) and situation (SC_q) clusters.

	SC_1	SC_2	
PC_1	+0.01	-0.05	48 pers.
PC_2	+0.31	-1.33	20 pers.
PC_3	-0.20	+0.86	34 pers.
	13 sit.	3 sit.	

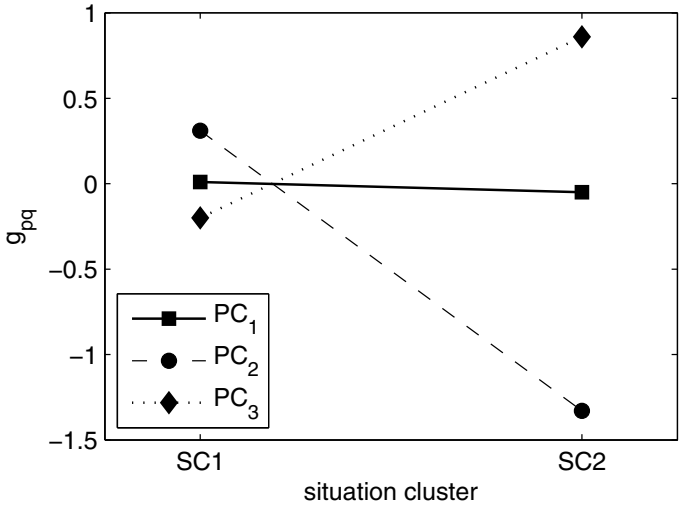

Figure 1. Visual display of the block-specific interaction terms $g_{p,q}$ for all pairs of person (PC_p) and situation (SC_q) clusters.

which one feels capable to empathize with others). The classification accuracy of predicting W using these two predictor variables amounts to 54%, which is significantly more than can be expected by chance. The estimated regression coefficients $\hat{\theta}_{cl}$ for this multinomial regression are presented in Table 2.

Combining the results of the analyses discussed above, it appears from Table 2 and Figure 1 that a smaller amount of satisfaction when helped by others (which holds for PC_2 persons, that is, persons for which $W_i = 2$) leads to less helping behavior in more ambiguous situations (i.e., the situations of SC_2) than can be expected from an additive effect of the persons and situations. In contrast, more satisfaction when helped by others combined with feeling less capable of empathizing with others (which holds for PC_3 persons, that is, persons for which $W_i = 3$) leads to more helping behav-

Table 2. Estimated regression coefficients $\hat{\theta}_{cl}$ for multinomial regression of person cluster membership W on Z_1 (i.e., the extent to which one feels satisfied when being helped by others) and Z_2 (i.e., the extent to which one feels capable to empathize with others).

Group discrimination	Z_1	Z_2
$W_i = 2$ versus $W_i = 1$	-.17	.06
$W_i = 3$ versus $W_i = 1$	.19	-.30

ior in ambiguous situations than can be expected on the basis of an additive effect of the persons and situations. At first sight, the lower self-reported level of empathy of PC_3 persons may seem counterintuitive. Yet, it can be explained in that in more ambiguous situations, the PC_3 persons, who feel the least capable of empathizing with others, will let their intention to help be driven mainly by the fact that they themselves would feel highly satisfied when helped by others.

6. Discussion and Conclusion

6.1 Relation to Other Biclustering Methods Involving Interaction Concepts

After the paper by Bock (1980), various biclustering methods have been proposed that are based on interaction concepts. However, when comparing these methods it should be kept in mind that the approaches may refer to at least three different types of deviations from additivity, i.e.:
– block-specific deviations from additivity $g_{p,q} = \bar{d}_{R_p,C_q} - \bar{d}_{R_p,\cdot} - \bar{d}_{\cdot,C_q} + \bar{d}_{\cdot,\cdot}$ from (1) used in the maximum interaction approach
– individual overall deviations from additivity $d^*_{i,j} = d_{i,j} - \bar{d}_{i,\cdot} - \bar{d}_{\cdot,j} + \bar{d}_{\cdot,\cdot}$ from (12)
– and individual bicluster-specific deviations from additivity

$$s^{(p,q)}_{i,j} := d_{i,j} - \bar{d}_{i,C_q} - \bar{d}_{R_p,j} + \bar{d}_{R_p,C_q} \qquad i \in R_p, j \in C_q. \tag{15}$$

Each of these may be useful for handling specific research questions.

A method that involves the individual bicluster-specific deviations from additivity (15) was proposed by Cheng and Church (2000) and is now one of the most cited methods in biclustering of microarray data. Specifically, Cheng and Church look for one or a few (possibly overlapping) biclusters $R_p \times C_q$ (with a row cluster $R_p \subset \mathcal{X}$ and a column cluster $C_q \subset \mathcal{Y}$) of maximal size and that are such that within $R_p \times C_q$ the mean-squared value of the individual bicluster-specific deviations from additivity $s^{(p,q)}_{i,j}$ is smaller than some pre-specified threshold ϵ, say. After such a bicluster has

been found, the algorithm replaces the entries of this bicluster by random draws of a uniform distribution and may proceed in the same way to find additional biclusters in a stepwise fashion. In a related method, Cho et al. (2004) look for a two-mode partitioning $\mathcal{R} \times \mathcal{C}$ with a minimum overall sum $H := \sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{i \in R_p} \sum_{j \in C_q} |s_{i,j}^{(p,q)}|^2$. A discussion of these and closely related biclustering methods can be found in Madeira and Oliveira (2004) and Tanay, Sharan, and Shamir (2005).

Both approaches differ from our maximal interaction clustering approach in some important respects. First, both approaches look for biclusters with (absolutely) *small* or *minimum* sums of squared interaction-type values $s_{i,j}^{(p,q)}$ while our approach looks for bipartitions with a *maximum* sum of squared block-specific interaction values $g_{p,q}$. Insofar, both Cheng and Church as well as Cho et al. look for biclusters with a negligible within-bicluster interaction while our approach generates biclusters such that row and column clusters show *high* between-bicluster interaction. It must be noted that minimizing the within-bicluster sums of squares of the individual bicluster-specific deviations from additivity does not imply that the overall between-bicluster sums of squares of block-specific deviations from additivity is maximized.

This can be illustrated insightfully by a small example. Consider the following data matrix

$$\mathbf{D} = \begin{bmatrix} 21 & 22 & 20 & 20 \\ 1 & 2 & 0 & 0 \\ 0 & 20 & 1 & 21 \\ 0 & 20 & 2 & 22 \end{bmatrix}.$$

The overall individual deviations from additivity sum of squares T is equal to 390.0. Assuming two row clusters and two column clusters, the optimal solution according to the criterion H by Cho et al. includes row clusters $R_1 = \{1, 2\}$ and $R_2 = \{3, 4\}$, and column clusters $C_1 = \{1, 2\}$ and $C_2 = \{3, 4\}$. Note that according to the Cho et al. criterion this is indeed the optimal biclustering as it yields $H = 0$, that is, within each of the four biclusters $R_p \times C_q$ ($p = 1, 2, q = 1, 2$) there are no interactions and only row and/or column main effects. However, this very same solution is *not* optimal according to interaction criterion (2). Specifically, the latter criterion takes a value of 9.0 for this solution. If, however, the column clusters are defined as $C_1 = \{1, 3\}$ and $C_2 = \{2, 4\}$ (while keeping the same row clusters), the value of interaction criterion (2) increases to 380.25, explaining 97.5% of the overall individual deviations from additivity sum of squares T.

Second, in the maximal interaction approach, after having removed the overall main effects of rows and columns, *homogeneous* biclusters are retained (see Remark 2 in Section 4.1), whereas Cheng and Church as well

as Cho et al. consider *heterogeneous* biclusters with bicluster-specific main effects of rows and columns.

Third, although the biclusterings resulting from the maximal interaction and Cho et al. approaches both capture the total row by column interaction sum of squares, in the maximal interaction clustering approach this interaction is represented by the between-bicluster differences with regard to the bicluster centroids only, whereas in the Cho et al. approach it is represented by the between-bicluster differences with regard to the bicluster centroids *plus* the between-bicluster differences with regard to the bicluster-specific main effects.

Another interaction-related biclustering method is that of Corsten and Denis (1990), who proposed a method for"identifying simultaneously groups of unstructured rows and groups of unstructured columns (...) such that the interaction between row and column factors is due only to interactions between those groups" (p. 207). This approach is based on an agglomerative hierarchical clustering procedure in each step of which either two rows (or row classes), or two columns (or column classes) are merged into one row or column class, respectively, based on proximity measures between all pairs of possibly merged rows and all pairs of possibly merged columns. This proximity measure is defined by the mean square for interaction in the data subset consisting only of the two rows or the two columns concerned. This method differs from our approach, among other things, in that it is a procedural clustering approach and as such does not involve an overall objective function (Van Mechelen, Bock, and De Boeck, 2004). Specifically, the stepwise approach considers only local interactions (i.e., calculated within the subset of data considered in each step) and therefore is at best indirectly related to the overall interaction criterion (2). To illustrate the difference, we reanalyzed the genotype by location data reported in Corsten and Denis (1990) by maximal interaction clustering (using the same software and algorithmic specifications as discussed in Section 5) and assuming 4 row clusters and 3 column clusters, since these are the numbers of row and column clusters of the solution reported by Corsten and Denis. The value of criterion (2) is equal to 933.6 for the biclustering solution we obtained by means of maximal interaction clustering, whereas it is equal to 555.3 for the solution reported by Corsten and Denis. This is a rather large difference considering the fact that the total row by column interaction sum of squares for this data set equals $T = 2108.4$.

Finally, a major difference between the interaction-related biclustering methods discussed above and maximal interaction biclustering is that this latter one can be justified as a classification likelihood estimation of probabilistic model (3) while such a theoretical basis is missing for the other approaches discussed above. Note that a purely additive model of the type

$d_{i,j} = \mu + \alpha_p + \beta_q + \epsilon_{i,j}$ for $i \in R_p, j \in C_q$ in analogy to (3) or (10) would lead, via a classification likelihood, to separately clustering the rows and columns of $\mathbf{D}$ according to the classical least-squares k-means clustering criterion (see Bock, 1968, pp. 40–43).

6.2 Main Effects vs. Interaction

Various two-mode clustering methods are able to capture to some extent row by column interactions. For instance, classical least-squares two-mode partitioning (see Section 3.2), when applied to an arbitrary observed data matrix $\mathbf{D}$, will typically yield a block centroid matrix $\hat{\mathbf{M}} = (\hat{m}_{p,q})$ with entries $\hat{m}_{p,q}$ that are not equal to a sum of row and column main effects in $\hat{\mathbf{M}}$ and that also include interaction effects. However, when for instance the row main effect in the data matrix $\mathbf{D}$ is very large, the resulting partition of the rows will largely be such that it captures this main effect as well as possible, and likewise in the case of a large column main effect. Indeed, in applications of classical least-squares two-mode partitioning, it is observed frequently that the obtained solution is mostly dominated by main effects of the row and/or column clustering(s). As discussed in the Introduction section, depending on the substantive-theoretical research question at hand, this may be undesirable. Maximal interaction two-mode clustering (which eliminates the main effects from the very beginning) may then be preferred because it focuses only on the row by column interaction.

From the discussion so far, it may appear that in the context of biclustering interest in structuring the row by column interaction is eventually at odds with interest in clustering the row and column main effects since, clearly, a biclustering $\mathcal{R} \times \mathcal{C}$ of the main effects will not necessarily be identical or even similar to a biclustering $\mathcal{R}' \times \mathcal{C}'$ of the interaction only. Interestingly, however, the maximal interaction two-mode clustering approach can be complemented by analyses that exclusively focus on the main effects of the rows and columns of $\mathbf{D}$. In particular, such analyses could be done by applying either least-squares one-mode partitioning to the row (resp. column) means of $\mathbf{D}$, or by applying classical least-squares two-mode partitioning to the matrix $\tilde{\mathbf{D}} = (\tilde{d}_{i,j})$, where $\tilde{d}_{i,j} = d_{i,j} - d^*_{i,j}$ denotes the deviations between the observed data and their double-centered counterparts. From a substantive point of view, such a combined approach may be sensible because it allows for distinct models describing the main effects structure on the one hand and the interaction effect structure on the other hand. This may allow the user to distinguish the underlying mechanisms that drive the main effects from the underlying mechanisms that drive the interaction.

Example: To illustrate the case where the clustering structure of the main effects and the clustering structure of the interaction are not captured by the

Table 3. A 6×4 data matrix $\mathbf{D} = (d_{i,j})$, and corresponding individual row main effects α_i, individual column main effects β_j and individual row by column interaction terms $\gamma_{i,j}$, such that $d_{i,j} = \mu + \alpha_i + \beta_j + \gamma_{i,j}$ with $\mu = 5$.

	$d_{i,j}$				α_i	$\gamma_{i,j}$			
	$j=1$	$j=2$	$j=3$	$j=4$		$j=1$	$j=2$	$j=3$	$j=4$
$i=1$	7	-1	5	-3	-3	4	-4	4	-4
$i=2$	7	-1	5	-3	-3	4	-4	4	-4
$i=3$	10	2	8	0	0	4	-4	4	-4
$i=4$	2	10	0	8	0	-4	4	-4	4
$i=5$	5	13	3	11	3	-4	4	-4	4
$i=6$	5	13	3	11	3	-4	4	-4	4
β_j	1	1	-1	-1					

same bipartition, Table 3 shows a small example of an hypothetical 6×4 data matrix **D**. The entries of this data matrix were generated according to the formula $d_{i,j} = \mu + \alpha_i + \beta_j + \gamma_{i,j}$ with differently clustered main and interaction effects. Table 3 shows that the row main effect structure is induced by three row clusters $R_1 = \{1, 2\}$, $R_2 = \{3, 4\}$, and $R_3 = \{5, 6\}$ (see the α_i), and the column main effect structure by two column clusters $C_1 = \{1, 2\}$ and $C_2 = \{3, 4\}$ (see the β_j). In contrast, the structure of the interaction terms $\gamma_{i,j}$ is represented by a bipartition with two row clusters $\tilde{R}_1 = \{1, 2, 3\}$ and $\tilde{R}_2 = \{4, 5, 6\}$ and two column clusters $\tilde{C}_1 = \{1, 3\}$ and $\tilde{C}_2 = \{2, 4\}$. In fact, applying the maximal interaction criterion (2) to the data matrix **D** will reconstruct the latter interaction-based bipartition, while the result from classical least-squares two-mode partitioning using criterion (11) will typically miss this bipartition (and also the one related to the main effects).

6.3 Extension to More Than Two Categorical Predictor Variables

Extending maximal interaction two-mode clustering to more than two categorical predictor variables is straightforward, as long as the data have been collected with a fully factorial design (which implies that they can be arranged into an N-way N-mode array $\underline{\mathbf{D}}$). Perhaps even more so than in the two-mode case, a reduction of the number of elements of each mode is essential in case one wishes to understand the structural information that pertains to the interactions in large multiway data arrays $\underline{\mathbf{D}}$. This could be achieved in a similar way as discussed in this paper for the two-mode case. In particular, in the three-way case one should first triple-center the observed data array $\underline{\mathbf{D}}$ such that $d^*_{i,j,k} = d_{i,j,k} - \bar{d}_{i,\cdot,\cdot} - \bar{d}_{\cdot,j,\cdot} - \bar{d}_{\cdot,\cdot,k} + 2\bar{d}_{\cdot,\cdot,\cdot}$ (similarly to the definition of $d^*_{i,j}$ in Subsection 4.1), and subsequently apply least-

squares three-mode partitioning to the triple-centered data making use of standard three-mode partitioning algorithms (Kiers, 2004; Schepers, Van Mechelen, and Ceulemans, 2006). Further extensions to the N-way case with $N > 3$ are straightforward.

6.4 Conclusion

In this paper, we have shown that the interaction clustering criterion (2), proposed by Bock (1980), which focuses on the row cluster by column cluster interaction, can be justified in terms of a specific probabilistic ANOVA model (3) with individual main effect terms and block-specific interaction terms. In particular, we have shown that maximizing the classification likelihood for this probabilistic model is equivalent to maximizing Bock's interaction criterion (2). This result is useful because it facilitates comparisons of maximal interaction two-mode clustering to other clustering approaches and facilitates understanding the conditions under which the former approach is likely to be successful.

Secondly, we have shown that maximizing Bock's interaction clustering criterion (2) (i.e., maximizing the classification likelihood for model (3)) is equivalent to minimizing the classical least-squares two-mode partitioning criterion (11) when applied to the data matrix after double-centering. This latter result is useful because many good algorithms for classical least-squares two-mode partitioning already exist such that it is no longer necessary to develop a special one for the maximum interaction criterion (2).

References

BAIER, D., GAUL, W., and SCHADER, M. (1997), "Two-Mode Overlapping Clustering with Applications to Simultaneous Benefit Segmentation and Market Structuring", in *Classification and Knowledge Organization*, eds. R. Klar and O. Opitz, Berlin: Springer, pp. 557–566.

BANFIELD, J., and RAFTERY, A. (1993), "Model-Based Gaussian and Non-Gaussian Clustering, *Biometrics, 49*, 803–821.

BOCK, H-H. (1968), "Statistische Modelle für die Einfache und Doppelte Klassifikation von Normalverteilten Beobachtungen [Statistical Models for the One-Way and Two-Way Classification of Normally Distributed Observations], Ph. D. thesis, Albert-Ludwigs-Universität zu Freiburg, Germany.

BOCK, H-H. (1980), "Simultaneous Clustering of Objects and Variables", in *Analyse de Données et Informatique. Cours de la Commission des Communautés Européennes à Fontainebleau, 19-30 Mars 1979*, eds. R. Tomassone, M. Amirchhay, and D. Néel, Le Chesnay, France: Institut National de Recherche en Informatique et en Automatique (INRIA), pp. 187–203.

BOCK, H-H. (1996), "Probabilistic Models in Cluster Analysis", *Computational Statistics and Data Analysis, 23*, 5–28.

CARROLL, J., and ARABIE, P. (1980), "Multidimensional Scaling", *Annual Review of Psychology, 31*, 607–649.

CASPI, A., and MOFFITT, T. (2006), "Gene-Environment Interactions in Psychiatry: Joining Forces with Neuroscience", *Nature Reviews Neuroscience, 7*, 583–590.

CASTILLO, W., and TREJOS, J. (2002), "Two-Mode Partitioning: Review of Methods and Application of Tabu Search", in *Classification, Clustering, and Related Topics. Recent Advances and Applications. Studies in Classification, Data Analysis, and Knowledge Organization*, eds. K. Jajuga, A. Sokolowski, and H-H. Bock, Heidelberg, Germany: Springer-Verlag, pp. 43–51.

CEULEMANS, E., and KIERS, H. (2006), "Selecting Among Three-Mode Principal Component Models of Different Types and Complexities: A Numerical Convex Hull Based Method", *British Journal of Mathematical and Statistical Psychology, 59,* 133–150.

CHENG, Y., and CHURCH, G. (2000), "Biclustering of Expression Data", in *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology*, pp. 93–103.

CHO, H., DHILLON, I., GUAN, A., and SRA, S. (2004), "Minimum Sum-Squared Residue Co-Clustering of Gene Expression Data", in *Proceedings of the 4th SIAM International Conference on Knowledge Discovery and Data Mining*, pp. 124–125.

CORSTEN, L., and DENIS, J. (1990), "Structuring Interacion in Two-Way Tables by Clustering", *Biometrics, 46,* 207–215.

FORKMAN, J., and PIEPHO, H.-P. (2014), "Parametric Bootstrap Methods for Testing Multiplicative Terms in GGE and AMMI Models", *Biometrics, 70,* 639–647.

GABRIEL, K. (1971), "The Biplot Graphic Display of Matrices with Application to Principal Component Analysis", *Biometrika, 58*, 453–467.

GAUCH, H. (2006), "Statistical Analysis of Yield Trials by AMMI and GGE", *Crop Science, 46,* 1488–1500.

GAUCH, H., PIEPHO, H.-P., and ANNICCHIARICO, P. (2008), "Statistical Analysis of Yield Trials by AMMI and GGE: Further Considerations", *Crop Science, 48,* 866–889.

GAUL, W., and SCHADER, M. (1996), "A New Algorithm for Two-Mode Clustering", in *Data Analysis and Information Systems. Studies in Classification, Data Analysis, and Knowledge Organization*, eds. H-H. Bock and W. Polasek, Berlin, Germany: Springer, pp. 15–23.

GEISER, C., LITSON, K., BISHOP, J., KELLER, B., BURNS, G., SERVERA, M., and SHIFFMAN, S. (2015), "Analyzing Person, Situation and Person X Situation Interaction Effects: Latent State-Trait Models for the Combination of Random and Fixed Situations", *Psychological Methods, 20,* 165–192.

GOLLOB, H. (1968), "A Statistical Model Which Combines Features of Factor Analytic and Analysis of Variance Techniques", *Psychometrika, 33,* 73–115.

GOVAERT, G., and NADIF, M. (2013), *Co-Clustering*, Chichester, UK: Wiley.

GOWER, J., and HAND, D. (1996), *Biplots*, London, UK: Chapman & Hall.

HANSOHM, J. (2001), "Two-Mode Clustering with Genetic Algorithms", in *Classification, Automation, and New Media. Studies in Classification, Data Analysis, and Knowledge Organization*, eds. W. Gaul and G. Ritter, Berlin, Germany: Springer, pp. 87–93.

HUBERT, L., and ARABIE, P. (1985), "Comparing Partitions", *Journal of Classification, 2,* 193–218.

HUNTER, D. (2005), "Gene-Environment Interactions in Human Diseases", *Nature Reviews Genetics, 6,* 287–298.

IOVLEFF, S., and SINGH BHATIA, P. (2015), "blockcluster: Coclustering Package for Binary, Categorical, Contingency and Continuous Data-Sets", R package version 4.0.2, https://CRAN.R-project.org/package=blockcluster.

KIERS, H. (2004), "Clustering All Three Modes of Three-Mode Data: Computational Posibilities and Problems", in *Proceedings in Computational Statistics*, ed. J. Antoch, Heidelberg, Germany: Springer, pp. 303–313.

MADEIRA, S., and OLIVEIRA, A. (2004), "Biclustering Algorithms for Biological Data Analysis: A Survey", *IEEE Transactions on Computational Biology and Bioinformatics, 1,* 24–45.

MCLACHLAN, G. (1982), "The Classification and Mixture Maximum Likelihood Approaches to Cluster Analysis", in *Handbook of Statistics (Vol.2)*, eds. P.R. Krishnaiah and L.N. Kanal, Amsterdam: North-Holland, pp. 199–208.

MISCHEL, W., and SHODA, Y. (1995), "A Cognitive-Affective System Theory of Personality: Reconceptualizing Situations, Dispositions, Dynamics, and Invariance in Personality Structure", *Psychological Review, 102,* 246–268.

MISCHEL, W., and SHODA, Y. (1998), "Reconciling Processing Dynamics and Personality Dispositions", *Annual Review of Psychology, 49,* 229–258.

MOFFITT, T., CASPI, A., and RUTTER, M. (2006), "Measured Gene-Environment Interactions in Psychopathology: Concepts, Research Strategies, and Implications for Research, Intervention, and Public Understanding of Genetics", *Perspectives on Psychological Science, 1,* 5–27.

NATIONAL INSTITUTE OF ENVIRONMENTAL HEALTH SCIENCES (2016), "Gene-Environment Interaction", retrieved November 1, 2016 from http://www.niehs.nih.gov/health/topics/science/gene-env/.

PIEPHO, H.-P. (1997), "Analyzing Genotype-Environment Data by Mixed Models with Multiplicative Terms", *Biometrics, 53,* 761–766.

PIEPHO, H.-P. (1999), "Fitting a Regression Model for Genotype by Environment Data on Heading Dates in Grasses by Methods for Nonlinear Mixed Models", *Biometrics, 55,* 1120–1128.

QUINTIENS, G. (1999), "Een Interactionistische Benadering van Individuele Verschillen in Helpen en Laten Helpen [An Interactionist Approach to Individual Differences in Helping and Allowing to Help]", unpublished master's thesis, KULeuven, Belgium.

ROCCI, R., and VICHI, M. (2008), "Two-Mode Multi-Partitioning", *Computational Statistics and Data Analysis, 52,* 1984–2003.

SCHEPERS, J., CEULEMANS, E., and VAN MECHELEN, I. (2008), "Selecting Among Multi-Mode Partitioning Models of Different Complexities: A Comparison of Four Model Selection Criteria", *Journal of Classification, 25,* 67–85.

SCHEPERS, J., and HOFMANS, J. (2009), "TwoMP: A MATLAB Graphical User Interface for Two-Mode Partitioning", *Behavior Research Methods, 41,* 507–514.

SCHEPERS, J., VAN MECHELEN, I., and CEULEMANS, E. (2006), "Three-Mode Partitioning", *Computational Statistics and Data Analysis, 51,* 1623–1642.

SHAFII, B., and PRICE, W. (1998), "Analysis of Genotype-by-Environment Interaction Using the Additive Main Effects and Multiplicative Interaction Model and Stability Estimates, *Journal of Agricultural, Biological, and Environmental Statistics, 3,* 335–345.

SHODA, Y., WILSON, N., CHEN, J., GILMORE, A., and SMITH, R. (2013), "Cognitive-Affective Processing System Analysis of Intra-Individual Dynamics in Collaborative Therapeutic Assessment: Translating Basic Theory and Research into Clinical Applications", *Journal of Personality, 81,* 554–1568.

SHODA, Y., WILSON, N., WHITSETT, D., LEE-DUSSUD, J., and ZAYAS, V. (2015), "The Person as a Cognitive Affective Processing System: Quantitative Idiography as an Integral Component of Cumulative Science", in *APA Handbook of Personality and Social Psychology: Vol.4. Personality Processes and Individal Differences*, eds. M.

Mikulincer and P. Shaver, American Psychological Association APA, Washington, pp. 491–513.

STEINLEY, D. (2004), "Properties of the Hubert-Arabie Adjusted Rand Index", *Psychological Methods, 9,* 386–396.

TANAY, A., SHARAN, R., and SHAMIR, R. (2005), "Biclustering Algorithms: A Survey", in *Handbook of Computational Molecular Biology*, ed. S. Aluru, Boca Raton: Chapman and Hall/CRC.

VAN MECHELEN, I., BOCK, H-H., and DE BOECK, P. (2004), "Two-Mode Clustering Methods: A Structured Overview", *Statistical Methods in Medical Research, 13,* 363–394.

VAN ROSMALEN, J., GROENEN, P., TREJOS, J., and CASTILLO, W. (2009), "Optimization Strategies for Two-Mode Partitioning, *Journal of Classification, 26*, 155–181.

VICHI, M. (2001), "Double K-Means Clustering for Simultaneous Classification of Objects and Variables", in *Advances in Classification and Data Analysis*, eds. S. Borra, R. Rocci, M. Vichi, and M. Schader, Berlin Heidelberg: Springer, pages 43–52.

WILDERJANS, T., CEULEMANS, E., and MEERS, K. (2013), "CHull: A Generic Convex Hull Based Model Selection Method", *Behavior Research Methods, 45,* 1–15.